CR-SAM: Curvature Regularized Sharpness-Aware Minimization

Tao Wu[1]; Tie Luo[1*] ; Donal C. Wunsch II[2]
Missouri University of Science and Technology
[1]Department of Computer Science, [2]Department of Electrical and Computer Engineering

BACKGROUND

- **Sharpness-Aware Minimization (SAM)** directs the search for model parameters within flat regions by minimizing the maximum loss in the vicinity of weight $\boldsymbol{w}$ within a radius ρ

$$\min_{\boldsymbol{w}} L^{\mathrm{SAM}}(\boldsymbol{w}) \quad \text{where} \quad L^{\mathrm{SAM}}(\boldsymbol{w}) = \max_{\|\boldsymbol{v}\|_2 \le 1} L_{\mathcal{S}}(\boldsymbol{w} + \rho \boldsymbol{v})$$

- SAM employs one-step gradient ascent to approximate the inner maximization

$$\nabla L^{\mathrm{SAM}}(\boldsymbol{w}) \approx \nabla L_{\mathcal{S}}\left(\boldsymbol{w} + \rho \frac{\nabla L_{\mathcal{S}}(\boldsymbol{w})}{\|\nabla L_{\mathcal{S}}(\boldsymbol{w})\|_2}\right)$$

PITFALLS IDENTIFIED (by us)

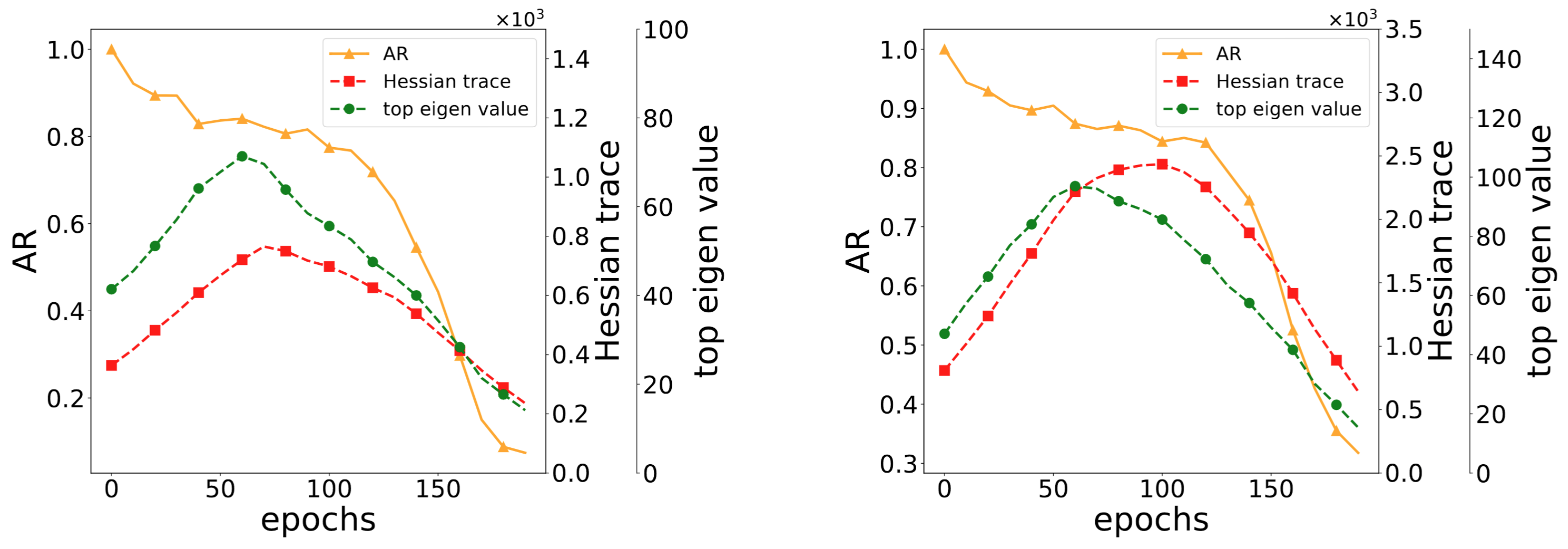
- **Deterioration of one-step sharpness approximation as training proceeds.**

  We define *approximation ratio* (AR):

$$\mathrm{AR} = \mathbb{E}_{(x,y)\sim D}\left[\frac{\ell\left(f\left(x;\boldsymbol{w}+\boldsymbol{\delta}\right),y\right) - \ell\left(f\left(x;\boldsymbol{w}\right),y\right)}{\ell\left(f\left(x;\boldsymbol{w}+\boldsymbol{\delta}^{*}\right),y\right) - \ell\left(f\left(x;\boldsymbol{w}\right),y\right)}\right]$$

  where $\boldsymbol{\delta}$ represents *one-step* gradient ascent perturbation, and $\boldsymbol{\delta}^*$ denotes the optimal perturbation which is calculated with 20-step gradient ascent.

- **Conventional Hessian fails to capture true loss landscape curvature**



(a) ResNet-18 on CIFAR10

(b) ResNet-18 on CIFAR100

Figure 1. Evolution of approximation ratio (AR), Hessian trace, and top eigenvalue of Hessian during SAM training on CIFAR10 and CIFAR100 datasets. The continuously decreasing AR indicates an enlarging curvature whereas both of the Hessian-based curvature metrics, Hessian trace and the top eigenvalue (which are expected to continuously increase) fail to capture the true curvature of model loss landscape.

RESOURCES AND CONTACT

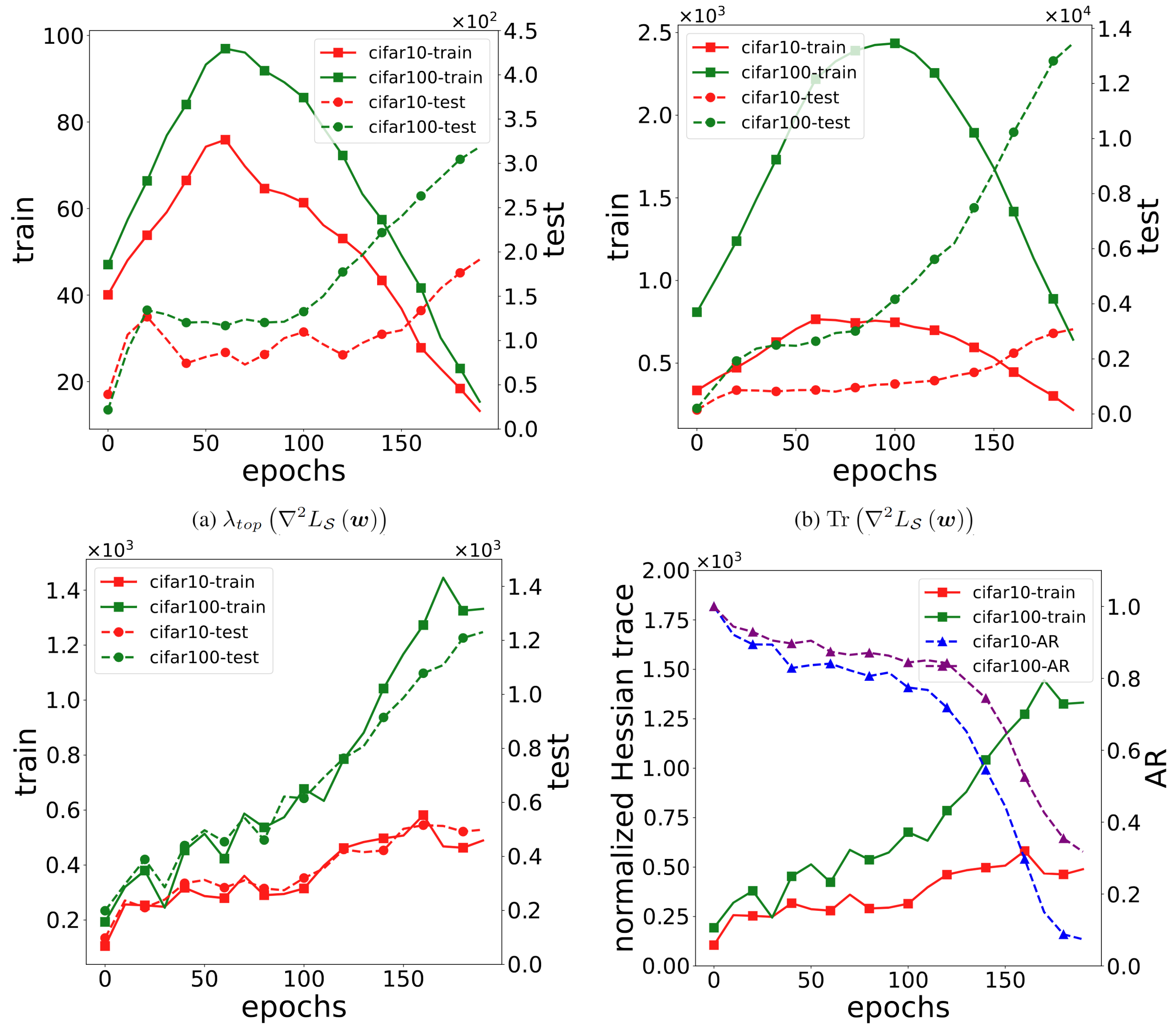
- **Paper**: https://arxiv.org/abs/2312.13555
- **Code**: https://github.com/TrustAIoT/CR-SAM
- **Contact**: wuta@mst.edu, tluo@mst.edu, dwunsch@mst.edu

METHODS

- **Propose a new metric for accurate curvature characterization, called *normalized Hessian trace*:**

$$\mathcal{C}(\boldsymbol{w}) = \frac{\mathrm{Tr}\left(\nabla^2 L_{\mathcal{S}}\left(\boldsymbol{w}\right)\right)}{\|\nabla L_{\mathcal{S}}(\boldsymbol{w})\|_2}$$

- **Characterizes curvature faithfully**
- **Consistent trends on both training and test sets**



(a) $\lambda_{top}\left(\nabla^2 L_{\mathcal{S}}\left(\boldsymbol{w}\right)\right)$

(b) $\mathrm{Tr}\left(\nabla^2 L_{\mathcal{S}}\left(\boldsymbol{w}\right)\right)$

(c) $\mathcal{C}(\boldsymbol{w})$

(d) $\mathcal{C}(\boldsymbol{w})$ *v.s.* AR

Figure 2. The normalized Hessian trace (d) indicates a growing curvature throughout training which explains the phenomenon of decreasing AR (approximation ratio). In addition, it behaves consistently on both train and test sets (**c**; as opposed to a & b which are conventional Hessian).

- **Curvature Regularized Sharpness-Aware Minimization (CR-SAM)**

$$R_c(\boldsymbol{w}) = \alpha \log \mathrm{Tr}\left(\nabla^2 L_{\mathcal{S}}\left(\boldsymbol{w}\right)\right) + \beta \log \|\nabla L_{\mathcal{S}}(\boldsymbol{w})\|_2$$

$$\min_{\boldsymbol{w}} L^{\text{CR-SAM}}(\boldsymbol{w})$$

$$\text{where } L^{\text{CR-SAM}}(\boldsymbol{w}) = L^{\mathrm{SAM}}(\boldsymbol{w}) + R_c(\boldsymbol{w})$$

Solving it with *finite difference method*:

$$R_c(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{v}\sim N(0,I)}\Big[\alpha \log\left(L_{\mathcal{S}}(\boldsymbol{w}+\rho\boldsymbol{v}) + L_{\mathcal{S}}(\boldsymbol{w}-\rho\boldsymbol{v}) - 2L_{\mathcal{S}}(\boldsymbol{w})\right) + \beta \log\left(L_{\mathcal{S}}(\boldsymbol{w}+\rho\boldsymbol{v}) - L_{\mathcal{S}}(\boldsymbol{w}-\rho\boldsymbol{v})\right)\Big]$$

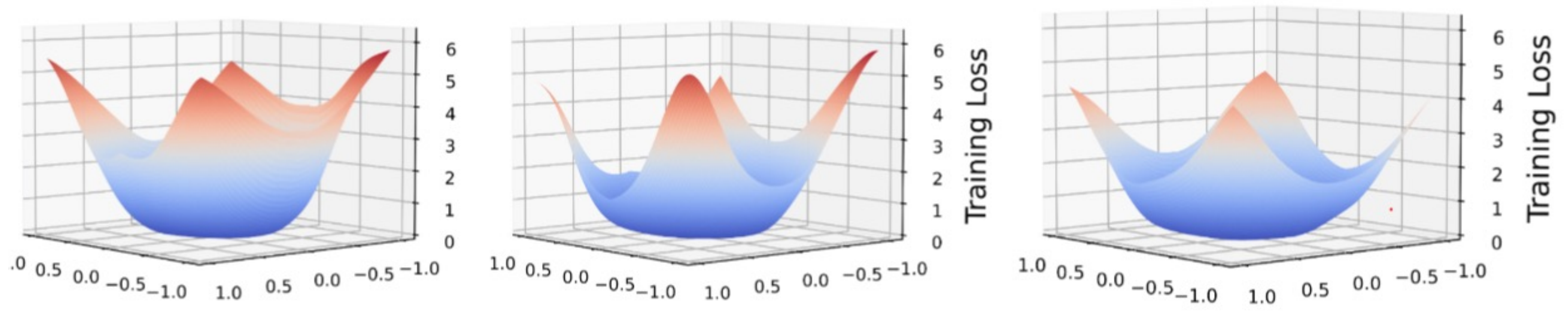
RESULTS

		CIFAR-10			CIFAR-100		
Model	Aug	SGD	SAM	CR-SAM	SGD	SAM	CR-SAM
ResNet-18	Basic	$95.29_{\pm 0.16}$	$96.46_{\pm 0.18}$	$\mathbf{96.95}_{\pm 0.13}$	$78.34_{\pm 0.22}$	$79.81_{\pm 0.18}$	$\mathbf{80.76}_{\pm 0.21}$
	Cutout	$95.96_{\pm 0.13}$	$96.55_{\pm 0.15}$	$\mathbf{97.01}_{\pm 0.21}$	$79.23_{\pm 0.13}$	$80.15_{\pm 0.17}$	$\mathbf{81.26}_{\pm 0.19}$
	AA	$96.33_{\pm 0.15}$	$96.75_{\pm 0.18}$	$\mathbf{97.27}_{\pm 0.12}$	$79.05_{\pm 0.17}$	$81.26_{\pm 0.21}$	$\mathbf{82.11}_{\pm 0.22}$
ResNet-101	Basic	$96.35_{\pm 0.12}$	$96.51_{\pm 0.16}$	$\mathbf{97.14}_{\pm 0.11}$	$80.54_{\pm 0.13}$	$82.11_{\pm 0.12}$	$\mathbf{83.03}_{\pm 0.17}$
	Cutout	$96.56_{\pm 0.18}$	$96.95_{\pm 0.13}$	$\mathbf{97.51}_{\pm 0.24}$	$81.26_{\pm 0.21}$	$82.39_{\pm 0.27}$	$\mathbf{83.46}_{\pm 0.16}$
	AA	$96.78_{\pm 0.14}$	$97.11_{\pm 0.16}$	$\mathbf{97.76}_{\pm 0.16}$	$81.83_{\pm 0.37}$	$83.25_{\pm 0.47}$	$\mathbf{84.19}_{\pm 0.23}$
WRN-28-10	Basic	$95.89_{\pm 0.21}$	$96.81_{\pm 0.26}$	$\mathbf{97.36}_{\pm 0.15}$	$81.84_{\pm 0.13}$	$83.15_{\pm 0.14}$	$\mathbf{84.45}_{\pm 0.09}$
	Cutout	$96.89_{\pm 0.07}$	$97.55_{\pm 0.16}$	$\mathbf{97.98}_{\pm 0.21}$	$81.96_{\pm 0.40}$	$83.47_{\pm 0.15}$	$\mathbf{84.48}_{\pm 0.13}$
	AA	$96.93_{\pm 0.12}$	$97.59_{\pm 0.06}$	$\mathbf{97.94}_{\pm 0.08}$	$82.16_{\pm 0.11}$	$83.69_{\pm 0.26}$	$\mathbf{84.74}_{\pm 0.21}$
PyramidNet-110	Basic	$96.27_{\pm 0.13}$	$97.34_{\pm 0.13}$	$\mathbf{97.89}_{\pm 0.08}$	$83.27_{\pm 0.12}$	$84.89_{\pm 0.09}$	$\mathbf{85.68}_{\pm 0.14}$
	Cutout	$96.79_{\pm 0.13}$	$97.61_{\pm 0.21}$	$\mathbf{98.08}_{\pm 0.11}$	$83.43_{\pm 0.21}$	$84.97_{\pm 0.17}$	$\mathbf{85.86}_{\pm 0.21}$
	AA	$96.97_{\pm 0.08}$	$97.81_{\pm 0.13}$	$\mathbf{98.26}_{\pm 0.11}$	$84.59_{\pm 0.08}$	$85.76_{\pm 0.23}$	$\mathbf{86.58}_{\pm 0.14}$

Table 1. Classification accuracy on CIFAR-10 and CIFAR-100 datasets.

Optimizer	$\|\nabla L_{\mathcal{S}}(\boldsymbol{w})\|_2$	$\mathrm{Tr}\left(\nabla^2 L_{\mathcal{S}}\left(\boldsymbol{w}\right)\right)$	$\mathcal{C}(\boldsymbol{w})$
SGD	19.97 ± 0.52	32673 ± 1497	1674 ± 78
SAM	11.51 ± 0.31	14176 ± 327	1193 ± 59
CR-SAM	**8.26** ± 0.19	**7968** ± 145	**884** ± 23

Table 2. Model geometry (characterized by 3 metrics) of ResNet-18 trained with SGD, SAM and CR-SAM. Values are computed on test set. It shows the CR-SAM optimizer achieves the minimal in all cases.



(a) SGD (b) SAM (c) CR-SAM

Figure 3. CR-SAM yields flatter loss landscape.

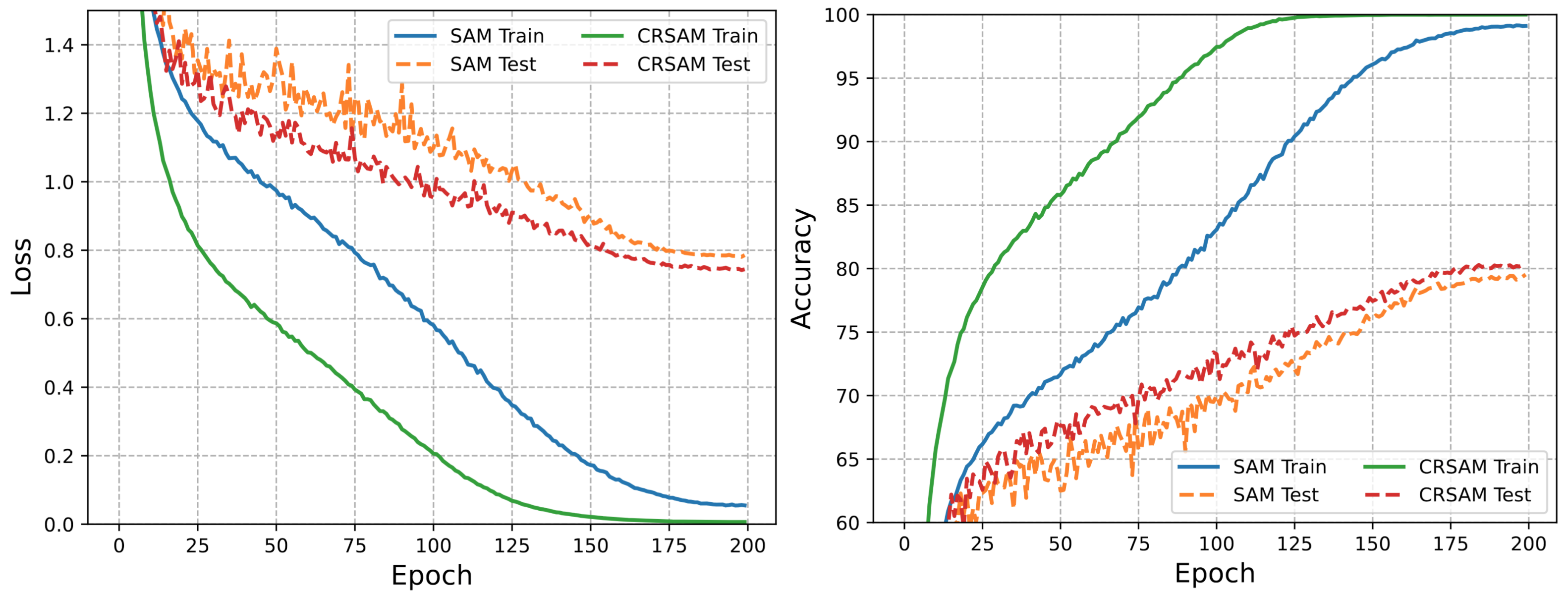


Figure 4. CR-SAM achieves much faster and stabler convergence. This can be explained by the fact that CR-SAM discourages excessive curvature and thus reduces optimization complexity, making local minima easier to reach.